CLAIMS

What is claimed is:

1. A method of policing resources in a computing utility facility, comprising:

intercepting a request for resources from an application admitted to access a pool of resources associated with the computing utility facility;

acquiring an entitlement profile associated with the application to determine if application is entitled to requested resources over a time period;

identifying an entitlement value and corresponding sliding window of the time period from the entitlement profile;

determining if the request for resources exceeds the entitlement value associated with the sliding window; and

indicating application entitlement to the request for resources in response to the determining if the request is excessive.

2. The method of claim 1 further comprising:

acquiring additional sliding windows and corresponding additional entitlement values to determine if the request for resources exceeds at least one entitlement value and sliding window combination; and

indicating that the application is not entitled to the requested resources when the request exceeds the entitlement value in at least one entitlement value and sliding window combination.

3. The method of claim 1 wherein the entitlement profile associated with the application describes the burstiness of the application over the time period.

4. The method of claim 1 wherein a burst loading factor associated with each sliding window corresponds to the burstiness of the application and identifies a portion of an aggregate entitlement to the resources available to fulfill the request.

5. The method of claim 4 wherein a larger burst loading factor is associated with more bursty applications that may need resources more rapidly compared with a smaller burst loading factor is associated with applications that may not need resources as rapidly.

6. The method of claim 1 wherein the entitlement value is derived from historical trace information collected while the application is using resources.

7. The method of claim 1 wherein the burst loading factor is derived from the historical trace information collected while the application is using resources.

8. The method of claim 3 wherein the resource usage is determined according to an estimated probability mass function.

9. The method of claim 4 wherein the estimated probability mass function further includes a confidence interval corresponding to a sample size used for determining the estimated probability mass function.

10. The method of claim 1 wherein the entitlement value operates as a metric for determining whether an application is entitled to the requested resources.

11. The method of claim 10 wherein the entitlement value for an application is proportional to the burstiness of the application in view of resource usage derived from historical trace data.

12. The method of claim 1 wherein determining if the request for resources exceeds the entitlement value further depends on a confidence interval associated with the entitlement value and the number of sample values used to identify the entitlement value.

13. The method of claim 1 wherein the indicating application entitlement includes throttling the requested resources when the application is not entitled to the additional resources.

14. The method of claim 1 wherein indicating application entitlement includes clawing back resources already allocated to the application when the application has exceeded a time limit for using the allocated resources.

15. An apparatus for policing resources in a computing utility facility, comprising:

a processor capable of executing instructions;

a memory containing instructions when executed cause the processor to intercept a request for resources from an application admitted to access a pool of resources associated with the computing utility facility, acquire an entitlement profile associated with the application to determine if application is entitled to requested resources over a time period, identify an entitlement value and corresponding sliding window of the time period from the entitlement profile, determine if the request for resources exceeds the entitlement value associated with the sliding window and indicate application entitlement to the request for resources in response to the determining if the request is excessive.

16. The apparatus of claim 15 further comprising instructions when executed that,

acquire additional sliding windows and corresponding additional entitlement values to determine if the request for resources exceeds at least one entitlement value and sliding

window combination and,

indicate that the application is not entitled to the requested resources when the request exceeds the entitlement value in at least one entitlement value and sliding window combination.

17. The apparatus of claim 15 wherein the entitlement profile associated with the application describes the burstiness of the application over the time period.

18. The apparatus of claim 15 wherein a burst loading factor associated with each sliding window corresponds to the burstiness of the application and identifies a portion of an aggregate entitlement to the resources available to fulfill the request.

19. The apparatus of claim 18 wherein a larger burst loading factor is associated with more bursty applications that may need resources more rapidly compared with a smaller burst loading factor is associated with applications that may not need resources as rapidly.

20. The apparatus of claim 15 wherein the entitlement value is derived from historical trace information collected while the application is using resources.

21. The apparatus of claim 15 wherein the burst loading factor is derived from the historical trace information collected while the application is using resources.

22. The apparatus of claim 17 wherein the resource usage is determined according to an estimated probability mass function.

23. The apparatus of claim 18 wherein the estimated probability mass function further includes a confidence interval corresponding to a sample size used for determining the estimated probability mass function.

24. The apparatus of claim 15 wherein the entitlement value operates as a metric for determining whether an application is entitled to the requested resources.

25. The apparatus of claim 24 wherein the entitlement value for an application is proportional to the burstiness of the application in view of resource usage derived from historical trace data.

26. The apparatus of claim 15 wherein determining if the request for resources exceeds the entitlement value further depends on a confidence interval associated with the entitlement value and the number of sample values used to identify the entitlement value.

27. The apparatus of claim 15 wherein the indicating application entitlement further includes instructions when executed that throttle the requested resources when the application is not entitled to the additional resources.

28. The apparatus of claim 15 wherein indicating application entitlement further includes instructions when executed that claw back resources already allocated to the application when the application has exceeded a time limit for using the allocated resources.

29. A computer program product for policing resources in a computing utility facility, comprising instructions operable to cause a programmable processor to:

intercept a request for resources from an application admitted to access a pool of resources associated with the computing utility facility;

acquire an entitlement profile associated with the application to determine if application is entitled to requested resources over a time period;

identify an entitlement value and corresponding sliding window of the time period from the entitlement profile;

determine if the request for resources exceeds the entitlement value associated with the sliding window; and

indicate application entitlement to the request for resources in response to the determining if the request is excessive.

30. An apparatus for policing resources in a computing utility facility, comprising:

means for intercepting a request for resources from an application admitted to access a pool of resources associated with the computing utility facility;

means for acquiring an entitlement profile associated with the application to determine if application is entitled to requested resources over a time period;

means for identifying an entitlement value and corresponding sliding window of the time period from the entitlement profile;

means for determining if the request for resources exceeds the entitlement value associated with the sliding window; and

means for indicating application entitlement to the request for resources in response to the determining if the request is excessive.